AI 特許紹介(72)

AI 特許を学ぶ！究める！

～Switch Transformer 特許～

2025 年 1 月 10 日

河野特許事務所

所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要

特許権者　Google

出願日　2023 年 7 月 7 日

登録日　2024 年 9 月 17 日

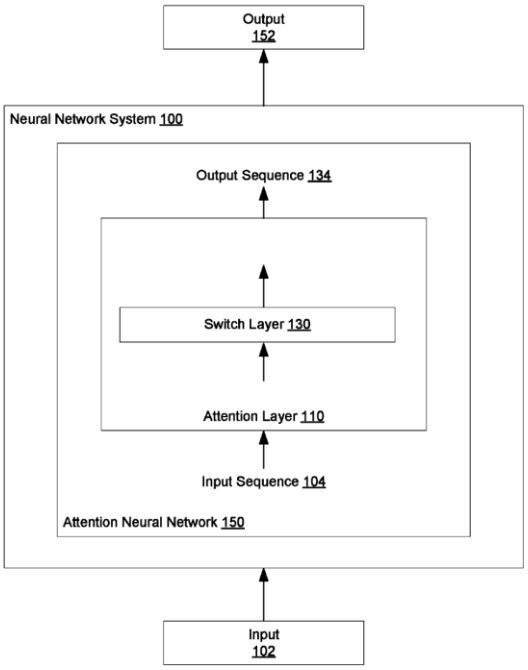
登録番号 US12093829

発明の名称　スイッチ層を備えたニューラルネットワーク

829 特許は、スイッチ層内の複数のエキスパートニューラルネットワークの各ルーティングスコアに応じてエキスパートニューラルネットワークを選択し最終出力を得ることで、計算コストを低減することが可能となる Switch Transformer 技術に関する。

2.特許内容の説明

図１は、ニューラルネットワークシステム 100 の一例を示す。



ニューラルネットワークシステム 100 は、入力 102 を受信し、入力 102 に対して機械学習タスクを実行して出力 152 を生成する。ニューラルネットワークシステム 100 には、複数のアテンション層 110 を含むアテンションニューラルネットワーク 150 が含まれている。

各アテンション層 110 は、入力シーケンス 104 に対して動作し、対応する出力シーケンス 134 を生成する。入力シーケンス 104 から出力シーケンス 134 を生成するために、各アテンション層 110 には、アテンションサブ層とフィードフォワードサブ層が含まれる。アテンションサブ層は、層 110 の入力シーケンス 104 を受け取り、層の入力シーケンスにアテンションメカニズムを適用して、アテンション入力シーケンスを生成する。

アテンションニューラルネットワーク内のアテンション層 110 には、スイッチ層 130 が含まれる。スイッチ層 130 は、学習済みルーティング機能と複数のエキスパートニューラルネットワークを含む層である。スイッチ層 130 によって処理される各入力に対して、スイッチ層 130 は、学習済みルーティング機能を使用して、層内の複数のエキスパートニューラルネットワークから 1 つのエキスパートニューラルネットワークを選択する。したがって、スイッチ層へのさまざまな入力に対して、どのエキスパートが単一のエキスパートとして選択されるかは変わる可能性があるが、特定の入力に対しては単一のエキスパートのみが選択される。

つまり、条件付き計算を採用する他のシステムとは異なり、スイッチニューラルネットワーク層では、特定の入力に対して複数のエキスパートが選択される可能性はなく、受信したすべての入力に対して単一のエキスパートのみが選択される必要がある。これにより、トレーニング後の推論の実行やトレーニング中のニューラルネットワークのフォワードパスに必要な FLOPS の数を増やすことなく、パラメータの数が大幅に増加し、ニューラルネットワーク 150 の計算能力が向上する。

次に、スイッチ層 130 は、選択されたエキスパートニューラルネットワークのみを使用して層入力を処理し、層入力の初期スイッチ層出力を生成し、選択されたエキスパートニューラルネットワークのルーティングスコアと、選択されたエキスパートニューラルネットワークによって生成された初期スイッチ層出力の積を計算することによって、スイッチ層の最終スイッチ層出力を生成する。

図 3A は、スイッチ層 330 を含むアテンションニューラルネットワーク層の例を示す。

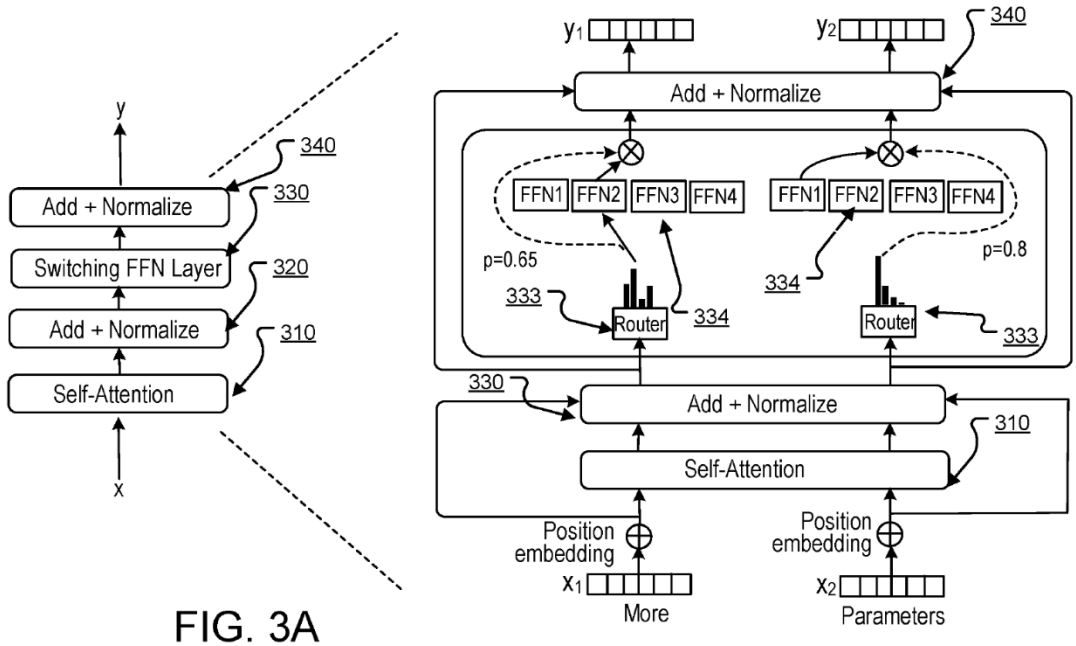


FIG. 3A

層には、層の入力シーケンスにアテンションメカニズム（図 3A の例では、セルフアテンション）を適用するアテンションサブ層 310 と、アテンション入力シーケンスを生成する「add & norm」操作 320 が含まる。「add & norm」操作 330 には、残差接続と、それに続く層正規化操作が含まれる。

アテンション層は、アテンション入力シーケンスを処理して、アテンション入力シーケンス内の各アテンション層入力のそれぞれの出力を含むアテンション層の出力シーケンスを生成する。

スイッチ層 330 は、アテンション入力シーケンスの各位置を個別に、つまり位置ごとに操作するように構成されている。特に、各入力位置について、スイッチ層 330 は、入力位置でアテンション層入力を受信し、入力位置でアテンション層入力に一連の変換を適用して、入力位置の出力を生成する。一般に、スイッチ層 330 は、特定の入力シーケンス内またはニューラルネットワークへの特定の入力バッチ内で、各スイッチ層入力の処理を並列に実行できる。これは、1 つのスイッチ層入力の計算が他のスイッチ層入力の計算とは独立しているためである。

次に、アテンション層は、初期出力に「add&norm」操作 340 を適用して、アテンション層の出力シーケンスを生成する。より具体的には、図 3A は、アテンションする入力シーケンスの 2 つの位置にある 2 つのトークン（1 つのトークン x1 は単語「more」に対応し、もう 1 つのトークン x3 は単語「parameters」に対応）に対するスイッチ層 330 の操作を示している。例のアテンション層はニューラルネットワークの最初のアテンション層であるため、アテンション層は、アテンションサブ層 310 と add＆norm 操作 330 を使用してトークンを処理する前に、まず各トークンに位置埋め込みを適用する。

スイッチ層 330 には、ルーティング機能 333 と 4 つのエキスパートニューラルネットワーク 334 が含まれる。各エキスパートニューラルネットワーク 334 は、フィードフォワードニューラルネットワーク（FFN)、たとえば、ReLU または GeLU アクティベーション関数を備えた完全に接続された層の多層（2 層または 3 層など）ニューラルネットワークにすることができる。

最初のトークンについては、スイッチ層 330 はルーティング機能 333 を最初のトークンに適用してエキスパートのスコアセットを生成する。最高スコアは、2 番目のエキスパート（FFN3）のスコア 0.65 である。これに基づいて、スイッチ層 330 は FFN3 のみを使用して 2 番目のトークンを処理し、FFN3 のエキスパート出力を生成し、次に、FFN3 のスコア(0.65)と FFN3 によって生成されたエキスパート出力の積を計算して、最初のトークンの最終出力を生成する。

2 番目のトークンについては、スイッチ層 330 はルーティング機能 333 を 2 番目のトークンに適用してエキスパートのスコアセットを生成する。最高スコアは、1 番目の

エキスパート（FFN1）のスコア 0.8 である。これに基づいて、スイッチ層 330 は、FFN1 のみを使用して 2 番目のトークンを処理し、FFN1 のエキスパート出力を生成し、次に、FFN1 のスコア(0.8)と FFN1 によって生成されたエキスパート出力の積を計算して、最初のトークンの最終出力を生成する。

3.クレーム

829 特許のクレーム 1 は以下の通りである。

1. ネットワーク入力に対して機械学習タスクを実行してネットワーク出力を生成するシステムにおいて、該システムは、1 台以上のコンピュータと、1 台以上のコンピュータによって実行されると 1 台以上のコンピュータに以下を実装させる命令を格納する 1 台以上のストレージデバイスとを備え、

機械学習タスクを実行するように構成されたニューラルネットワークを有し、ニューラルネットワークは 1 つ以上のスイッチ層を含み、各スイッチ層は以下を備える、

(i) ルーティングパラメータを有するそれぞれの学習されたルーティング機能と、

(ii) それぞれのエキスパートパラメータセットを有する複数のエキスパートニューラルネットワークのそれぞれが、スイッチ層のスイッチ層入力を受信し、スイッチ層のスイッチ層入力をエキスパートニューラルネットワークのそれぞれのエキスパートパラメータセットに従って処理し、スイッチ層のそれぞれの初期スイッチ層出力を生成するように構成されており、各スイッチ層は以下の通り構成されており、

スイッチ層のスイッチ層入力を受信し、

スイッチ層内のそれぞれの学習済みルーティング関数を、それぞれの学習済みルーティング関数のルーティングパラメータの現在の値に従ってスイッチ層入力に適用し、スイッチ層内の複数のエキスパートニューラルネットワークのそれぞれについてそれぞれのルーティングスコアを含むスコア分布を生成し、

複数のエキスパートニューラルネットワークから、ルーティングスコアが最も高いエキスパートニューラルネットワークのみを選択し、

層入力の初期スイッチ層出力を生成するために、選択されたエキスパートニューラルネットワークのみを使用して、選択されたエキスパートニューラルネットワークのエキスパートパラメータの現在の値に従ってスイッチ層入力を処理し、

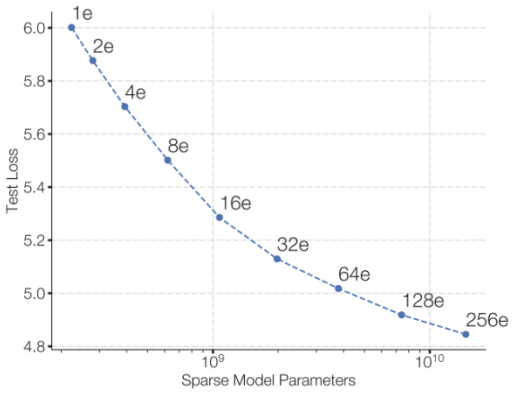
選択されたエキスパートニューラルネットワークのルーティングスコアと、選択されたエキスパートニューラルネットワークによって生成された初期スイッチ層出力との積を計算することを含む、スイッチ層の最終スイッチ層出力を生成する。

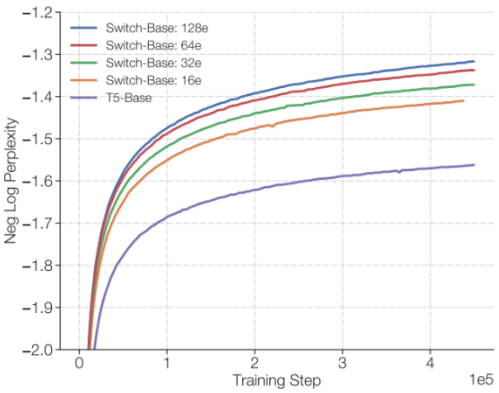
4. 本特許に関連する論文

本特許に関する論文“Switch Transformers: Scaling to Trillion Parameter Models

with Simple and Efficient Sparsity”[1]が、Google の William Fedus 氏らにより公表されている。

論文では上記アルゴリズムに加えて性能の向上を示すデータも示されている。下記図はスイッチトランスフォーマーのスケーリングプロパティを示すグラフである。





左のグラフは、 エキスパートの数をスケーリングしてパラメータが増加するにつれて、難しさで測定される品質の向上を測定している。左上の点は、2.23 億パラメータを持つ T5-Base モデルに対応する。左上から右下に移動すると、エキスパートの数を 2、4、8 と 2 倍に増やし、右下の点である 147 億パラメータを持つ 256 エキスパートモデルに到達する。すべてのモデルで同じ計算予算を使用しているにもかかわらず、エキスパートの数をスケーリングすると一貫した改善が見られる。

右のグラフは、エキスパートの数を網羅したステップごとの負の対数パープレキシティを示すグラフである。負の対数パープレキシティは、言語モデルの評価指標の一つであり、シーケンスの平均負の対数尤度を指数関数化したもので、値が低いほど、モデルの予測が正確であることを意味する。右のグラフに示すように密なベースラインは紫色の線で示され、Switch-Base モデルのサンプル効率が向上していることが理解できる。

以上

著者紹介
河野英仁
河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏

[1] William Fedus, et al. “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity” arXiv:2101.03961v3 [cs.LG] 16 Jun 2022

季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース、生成 AI ビジネスコース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。